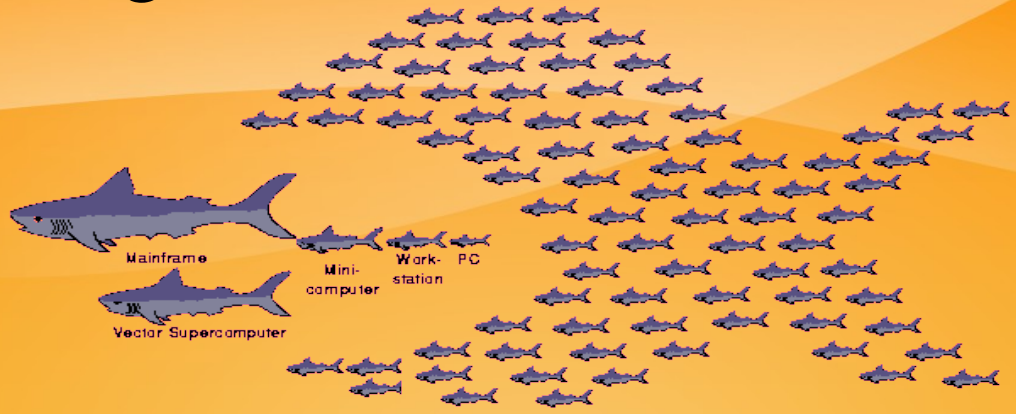
HA CLUSTERS

LINUXCABAL

26
abril
2014

¿Qué es un Clúster?



Definición: Un conjunto de cosas similares que ocurren juntas
http://www.merriam-webster.com/dictionary/cluster

Un cluster de computadores es un conjunto de computadoras interconectadas que trabajan en conjunto. La mayor parte de las veces se ve como si fuera un solo sistema.

Los componentes de un cluster son usualmente conectados uno a otro a a través de una Red de Area Local (LAN); por las altas velocidades, estos componentes son llamados nodos y cada uno de ellos ejecuta su propia instancia de un servicio.

Los cluster de computadoras han emergido como resultado de varios factores como la reducción de los costos de los microprocesadores, la alta velocidad de las redes y la aparición de software distribuido de alto rendimiento.

Los cluster son usualmente utilizados para mejorar el desempeño y la disponibilidad que pueda brindar una sola computadora,
Son una buena solución costo-beneficio comparados con grandes computadoras (velocidad y disponibilidad)

Los clusters de computadoras tienen un gran rango de aplicación, los cuales van desde cluster pequeños para pequeños negocios has grandes proyectos como las mas rápidas supercoputadoras.

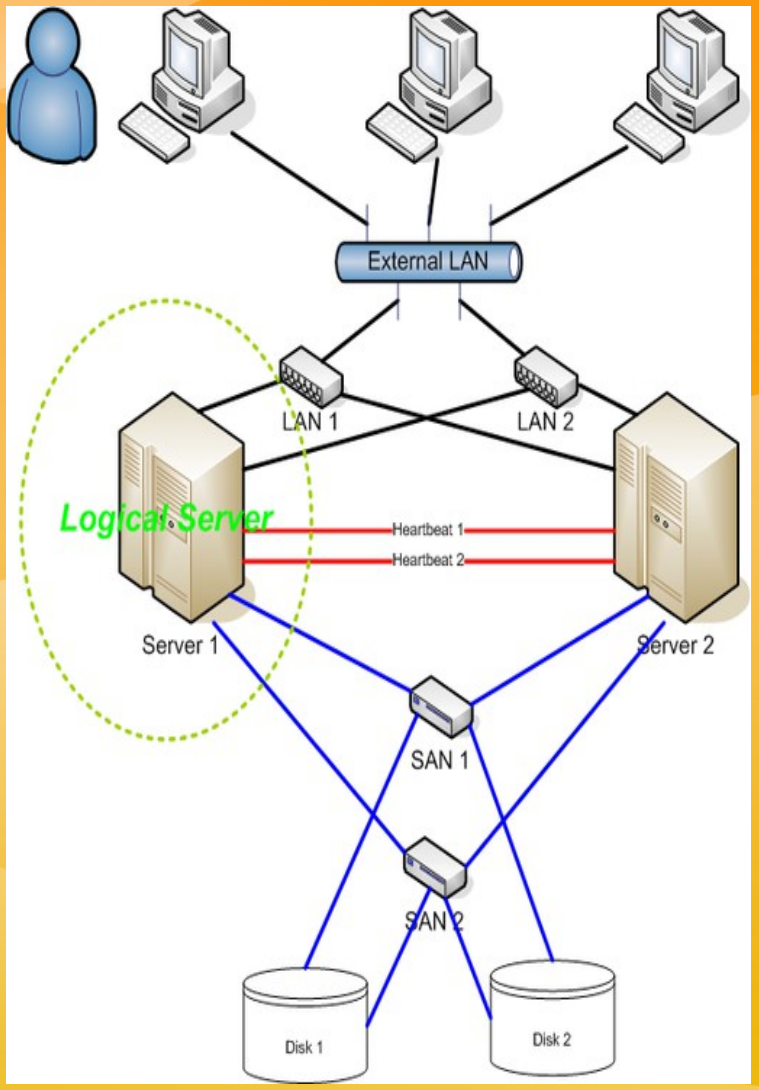
En informática, alta disponibilidad se refiere al sistema o componente que puede ser continuamente funcional por un periodo de tiempo largo. Disponibilidad se puede definir como “100% operativo” o “que nunca falla.”

Un amplio, pero difícil de conseguir, estándar de disponibilidad para un sistema o producto es conocido como “los cinco nueves” (99.999 por ciento) de disponibilidad.

Algunos expertos en disponibilidad enfatizan que para que un sistema sea considerado de alta disponibilidad, los componentes de un sistema deben de ser bien diseñados y completamente probados antes de ser puestos en producción.

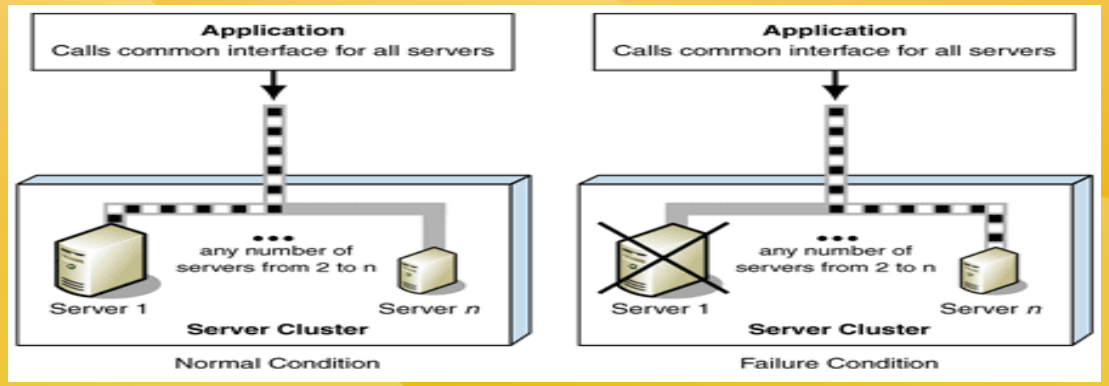
HA Clusters

- **Clusters de Alta-disponibilidad** (también conocidos como **HA clusters** o **failover clusters**) son un grupo de computadoras interconectadas que proveen de distintos servicios, estas computadoras serán utilizados para asegurar un tiempo de inactividad mínimo.
- HA clusters proveen acceso a un servicio incluso si algún componente del sistema falla. Sin ayuda de un HA cluster, si alguna aplicación fallara esta no estaría disponible hasta que se arreglara el problema en el servidor.
- HA clusters ha puesto solución al problema de fallas tanto de hardware como de software; incluso sin intervención de algún administrador, dicho proceso es conocido como **failover**.
- HA clusters son comúnmente utilizados para bases de datos criticas, servidores de archivos, servidores web y ciertas aplicaciones críticas.
- No solo la configuración de los nodos es redundante sino las conexiones de red y almacenamiento, esto para reducir cualquier falla en el sistema.
- Usualmente utilizan una red privada donde constantemente se monitorea el estatus de cada nodo, a esto se lo conoce como heartbeat.
- Haciendo a un lado todos los beneficios mencionados existe una peligrosa consideración a tomar en cuenta, el software debe de ser capaz de resolver el estado conocido como **split-brain.** Esta falla se presenta cuando existe un problema en la red y todos los enlaces se interrumpen (se caen) simultáneamente, sin embargo los nodos siguen funcionando. Si esto sucede cada nodo puede erróneamente decidir que los otros nodos se han caído e intentará iniciar los servicios los cuales los otros nodos aun siguen ejecutándose, esto ocasionará tener servicios duplicados y en el caso de tener almacenamiento compartido, la corrupción de datos en el shared storage.



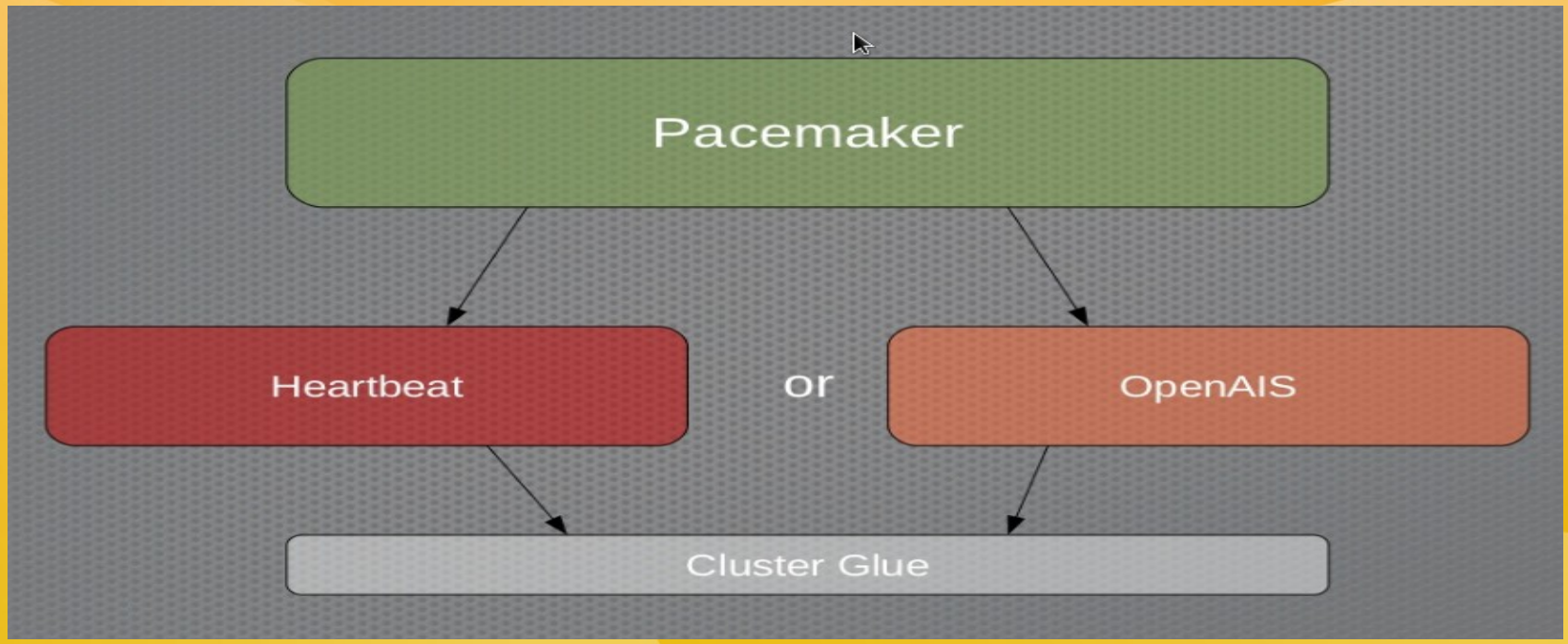
CONFIGURACIÓN DE NODOS

- El tamaño mas común para un HA clúster es de 2 nodos, porque es el mínimo requerido para proveer redundancia, aunque muchos clúster pueden tener muchos mas nodos, algunas veces hasta docenas de ellos. Algunas veces estas configuraciones se puede categorizar de la siguiente manera:
- **Activo/activo** — El trafico destinado para el nodo que falló es mandado al siguiente nodo o load balanced (balanceado) entre los nodos disponibles. Esto es usualmente posible cuando se utiliza una configuración homogénea entre los nodos.
- **Activo/pasivo** — Provee una redundancia completa para cada nodo, la cual es activada cuando el nodo primario falla. Esta configuración requiere mas hardware puesto que existe un nodo adicional por cada nodo activo.
- **N+1** — Provee un solo nodo el cual es puesto en línea para tomar el rol del nodo que falló. En el caso que se el clúster maneje diversos servicio el nodo deberá ser capaz de asumir el rol del nodo fallado, esto es será un nodo universal.
- **N+M** — En casos donde un clúster ofrece varios servicios tener un solo nodo en standby puede no ser suficiente. En tales casos se opta por tener mas un nodo (M) en standby para activarse en caso de ser necesario. La cantidad de nodos dependerá del costo-seguridad que se necesite.
- **N-to-1** — Permite al nodo que se activo convertirse en el nodo activo temporalmente, hasta que el nodo original sea restaurado y puesto en línea otra vez, esta configuración tiene la variante de que cuando el nodo activo regrese se tiene que hacer un failover hacia el nodo fallado para que se restaure el HA, esto es; el nodo pasivo se convierte en activo pero cuando el nodo original regrese a la normalidad este será fallado y regresaran los servicios a este una vez mas.
- **N-to-N** — Una combinación de activo/activo y N+M clúster, los clúster N to N redistribuyen los servicios, instancias y conexiones del nodo que falló entre los demás nodos activos, eliminando la necesidad de tener un nodo en stadby pero incrementando el costo y uso de recursos (red, almacenamiento, etc.)
- El término nodo virtual se refiere no a un solo nodo sino al servicio que el clúster brinda esto es, este responderá a una IP ofreciendo el servicio, no importando cual es el nodo que lo este ofreciendo.



HeartBeat

- Heartbeat es un demonio que provee un servicio de infraestructura de cluster (comunicación y asociación) a los clientes. Esto permite a los clientes del cluster conocer la presencia (o ausencia!) de procesos u otros equipos permitiendo el intercambio de mensajes entre ellos.
- Para que Heartbeat funcione adecuadamente el demonio de Heartbeat necesita ser trabajar de la mano con un administrador de recursos de cluster (**CRM cluster resource manager**) este tienes la función de iniciar y detener los servicios (IP addresses, web servers, etc.) que el cluster esta haciendo de alta disponibilidad. Pacemaker es el crm preferido para clusters basados en Heartbeat.



Pacemaker

Pacemaker es un open source HA-cluster resource manager. (CRM)

Como su nombre lo dice es el encargado de administrar recursos. Su trabajo consiste en proveer la máxima disponibilidad del clúster (recursos) detectando y restaurando fallos ocurridos en cada uno de los nodos.

Pacemaker mantiene la configuración de todos los recursos del cluster que serán administrados así también como la relación de los equipos y los recursos. A nivel de recursos se puede utilizar alguno de los dos sistemas de mensajes del clúster (OpenAIS o Heartbeat).

En caso de que un nodo falle, esta información sera inmediatamente transmitida a los nodos restantes del cluster.

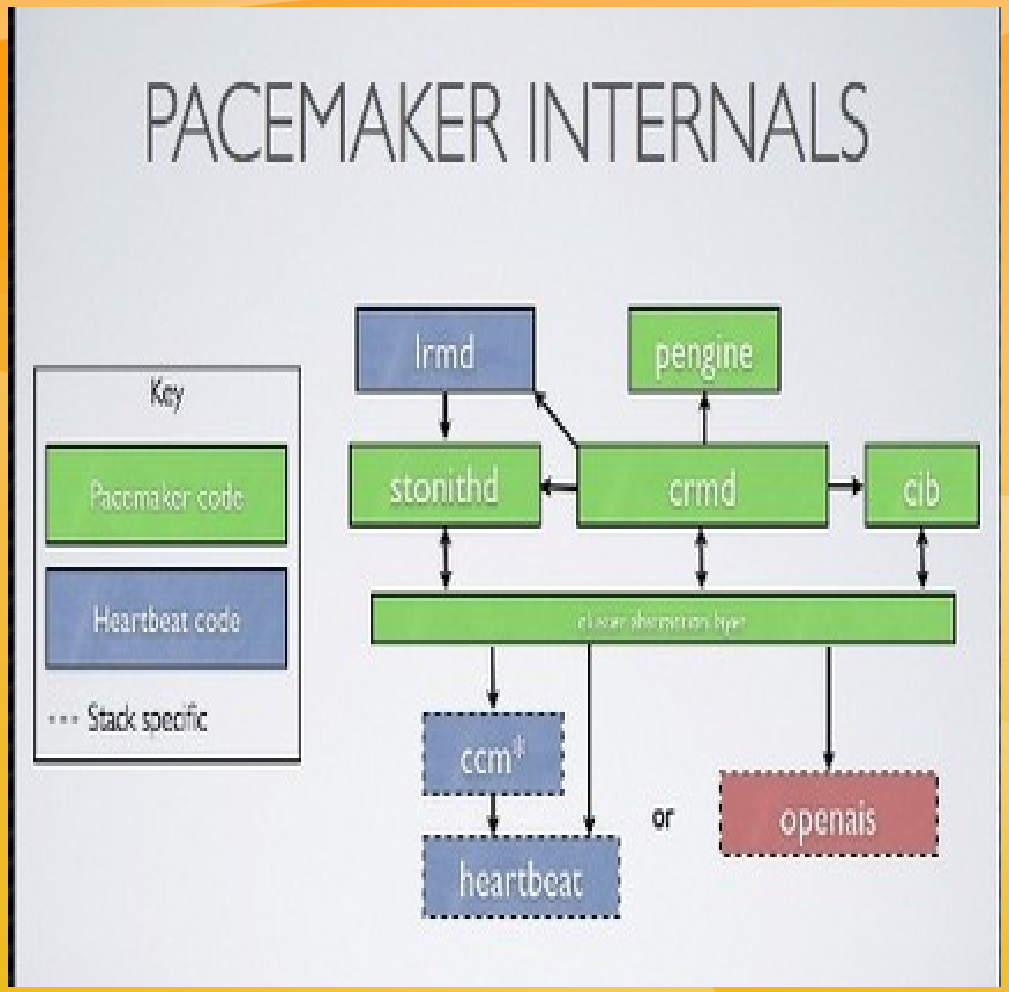
Corosync

Corosync es un open source Cluster Engine, esto es, un sistema de comunicación que permite que dos o mas nodos de un Linux cluster transfieran información entre ellos.

Corosync escucha constantemente en algún puerto, previamente configurado, donde los nodos que son parte del cluster estarán mandando información

El sistema de comunicación en Corosync permite que todos los nodos puedan conocer el estatus exacto de cada uno de los otros nodos en cualquier momento. En caso que uno de los nodos fallase esta información seria transmitida inmediatamente a todos los demás nodos para re-asignar recursos, reasignar tareas, etc.

Pacemaker Architeture



Sotnithd : The Heartbeat fencing subsystem.

Lrmd : Local Resource MAnagement Daemon. Interacts directly with resource agents (scripts).

pengine : Policy Engine. Computes the next state of the cluster based on the current state and the configuration.

cib : Cluste information Base. Contains definitions if all cluster options nodes, resouces, their relationships to one another and current status. Synchronizes updates to all cluster nodes.

crmd : Cluster Resource Management Daemon. Largely a message broker for the PEngine and LRM, it also elects a leader to co-rodinate the activities of the cluster.

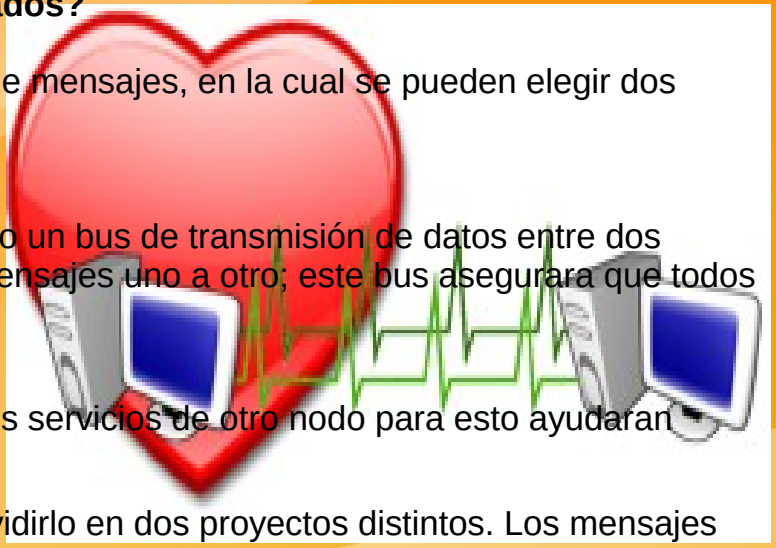
openais Messaging and membership layer

heartbeat Messaging layer, an alternative to OpenAIS

Ccm : Consensus Cluster Membership. The Heartbeat membership layer.

Pacemaker, Heartbeat, Corosync, WTH?

¿Qué son todos estos proyectos y como están relacionados?



El proyecto original de heartbeat fue divido en dos, la capa de mensajes, en la cual se pueden elegir dos opciones:

- **Heartbeat** - Capa de mensajes (Messaging layer)
  Se puede ejemplificar que a heartbeat y corosync como un bus de transmisión de datos entre dos nodos, de manera que cualquier nodo puede enviar mensajes uno a otro; este bus asegurara que todos reciban el mensaje.
- **Corosync** – Cluster Engine (Messaging layer)
  Pacemake no tiene la habilidad de comunicarse con los servicios de otro nodo para esto ayudaran corosync or heartbeat.

Corosync y OpenAIS eran lo mismo hasta que se decidió dividirlo en dos proyectos distintos. Los mensajes entre nodos y la capacidad para unir nodos al cluster es conocido hoy como corosync.

Y el sistema de administración de recursos:

- **Pacemaker** – Administrador de Recursos (Resource manager)
  Es el encargo de encender y detener servicios (la base de datos or el servidor de correo) este mantendrá la lógica del sistema ya que en el caso de accesos a un share compartido manejara que nodos mandaran información y cuales no esto es para evitar corromper la información.

Y sin olvidar los agentes.

- **Resource Agents** - Scripts para controlar varios servicios
  Son los encargados de ensenar a Pacemaker como manejar los distintos servicios que se quieran ofrecer como parte del cluster. Estos agentes (scripts) tendrán las instrucciones para realizar dicha tarea.

A pesar que Pacemaker solo necesita Corosync para funcionar existen algunas aplicaciones que necesitan a OpenAIS también para poder funcionar.

RECAPITULANDO

Clusters

- El término ***cluster*** se aplica a un conjunto de computadoras, construido utilizando componentes de hardware comunes y en la mayoría de los casos, software libre; los computadores se interconectan mediante alguna tecnología de red. El *cluster* puede estar conformado por nodos dedicados o por nodos no dedicados.
- Simplemente, un *cluster* es un grupo de múltiples computadoras unidas mediante una red de alta velocidad, de tal forma que el conjunto es visto como una única computadora.
- Para que un *cluster* funcione como tal, no basta solo con conectar entre sí las computadoras, sino que es necesario proveer un sistema de administración del *cluster*, el cual se encargue de interactuar con el usuario y los procesos que corren en él para optimizar su funcionamiento.

Beneficios de la Tecnología *Cluster*

- Las aplicaciones paralelas escalables requieren: buen rendimiento, baja latencia, comunicaciones que dispongan de gran ancho de banda, redes escalables y acceso rápido a archivos. Un *cluster* puede satisfacer estos requerimientos usando los recursos que tiene asociados a él.
- Los *clusters* ofrecen las siguientes características a un costo relativamente bajo:
  - Alto Rendimiento (High Performance).
  - Alta Disponibilidad (High Availability).
  - Alta Eficiencia (High Throughput).
  - Escalabilidad.
- La tecnología *cluster* permite a las organizaciones incrementar su capacidad de procesamiento usando tecnología estándar, tanto en componentes de hardware como de software que pueden adquirirse a un costo relativamente bajo.

RECURSOS

- http://clusterlabs.org/wiki/Pacemaker
- http://www.linux-ha.org

GRACIAS